An Integrated Framework for COVID and Pneumonia Disease Prediction using Optimized Deep Learning Models

By

Ms. S R Kalaiselvi

(19PHCSP004)

Supervisor

Dr. R. Vijayabhanu, Associate Professor

A Synopsis Submitted to

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore – 641 043

In partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

September 2024

An Integrated Framework for COVID and Pneumonia Disease Prediction Using Optimized Deep Learning Models

1. Introduction

The COVID-19 pandemic has highlighted the urgent need for advanced diagnostic tools to address a range of respiratory illnesses effectively. COVID and pneumonia are two significant diseases that impact respiratory health, though they manifest and progress differently. COVID, caused by the SARS-CoV-2 virus, can lead to a range of symptoms from mild to severe and often requires specific management strategies. On the other hand, pneumonia, which can be caused by various pathogens including bacteria and viruses, leads to inflammation of the lungs and affects breathing. Early forecast of these diseases is important for optimized clinical care and resource management, especially during an epidemic.

Data-mining methods are proving to be more and more helpful in the present day, mostly in the meadow of medical applications. By using Machine Learning (ML) methods, noteworthy information is extracted from the healthcare data that can be utilized to forecast diseases early, and thereby decreasing the risk of human life. In addition, the Deep Learning (DL) techniques are more helpful particularly in the field of health care in categorizing data from healthcare systems. Due to the requirements for extensive data pre-processing and feature selection, numerous ML and DL methods were developed for addressing COVID and Pneumonia diseases prediction. It is observed that the existing techniques and medical testing events with early analysis of infected patients, gathered samples require more time to provide optimal support to the patients. Several works are developed to find the better results to diminish virus extent. A well-known approach is unpredictable to the complexity of improving the trend of infections. Therefore, a novel technique is needed for reliable patient data prediction.

1.1 Problem statement

Distinguishing diseases from other respiratory infection with clinical features is relentlessly challenging and there is a need to provide insights to the specific scenario with the development of an efficient disease classification models at an early stage with minimum error , time and reduced space.

1.2 Objective

- To develop a disease prediction model for COVID and Pneumonia with improved performance using Machine Learning and Deep Learning techniques.

1.3 Secondary Objectives

- To enhance data preprocessing for high accuracy, techniques are proposed that effectively transform and standardize data to improve model performance.
- To improve the prediction accuracy and select the relevant features using proposed feature selection approach.
- To enhance the prediction model with high accuracy and reduce error rate using machine learning and deep learning techniques.

2. Review of Literature

Convolutional Neural Networks-Gated Recurrent Unit (CNN-GRU) based hybrid deep learning model developed by Anil Utku (2023) [1] to predict the disease spreading range. CNN was employed to extracting spatial features from the input data. However, disease prediction accuracy was not improved. VOC-DL prediction framework developed by Zhifang Liao *et al*., (2022) [2] to forecast the daily new confirmed cases. But, the other variants prediction was not carried out by designed framework. Advanced Hybrid Ensemble Gain ratio Feature Selection (AHEG-FS) model was designed by Syed Javeed Pasha *et al*., (2022) [3] to enhance the disease risk detection. But, the space complexity was not reduced. Samrat Kumar Dey *et al*., (2022) [4] introduced hybrid Chi-squared test and Mutual Information (Chi^2-MI) based feature selection model to identify the chronic disease. However, accuracy of feature selection was not improved with minimum error. Deep long short-term memory (LSTM ensemble model) was designed by Sourabh Shastri *et al*., (2021) [5] to forecast the COVID disease. But, it failed to present preventive measures for illness prediction cases.

Another prediction model called Deep Sequential Prediction Model (DSPM) was introduced by Devante Ayris *et al*., (2022) [6] to find the spread of infection with less error rates. However, the prediction performance rate was not enhanced. A novel hybrid feature representation method was designed by ChunyanAo *et al*., (2020) [7] for finding the antioxidant protein level. The

designed model was including hybridized random forest as well as hybrid features. But, the feature selection was not made with better precision. Yet another fusing matrix-based representation and convolutional neural networks (CNN) was proposed by Haolin Wang *et al*., (2020) [8] to extract the feature for medical decision making. But, the performance prediction results were not obtained. Deep contrastive mutual learning (DCML) was designed in Hongbin Zhang *et al*., (2022) [9] to identify disease. But, optimal performance was not attained. A Soft-Voting Ensemble Classifier was designed by Andrea Manconi *et al*., (2022) [10] for detecting the patients affected from disease. But, dataset complexity was not reduced.

Decision-making-basis federated learning network (DMFL_Net) was introduced by Malik h *et al*., (2022) [11] with less time. But, the accuracy was not improved. LSTM networks basis of the transfer learning approach was developed by Gautam Y (2022) [12] to forecast diseases. However, disease prediction time was not minimized. A novel cross-entropy basis loss function was presented by Anand Motwani *et al*., (2023) [13] to boost the CNN algorithm's convergence. However, the risk factors were not mitigated. The recurrent as well as CNN schemes were developed by Hanuman Verma *et al*., (2022) [14] to forecast the disease confirmed cases. But, error rate was not decreased. A depth wise separable convolution neural network (DWS-CNN) was developed by Le D *et al*., (2021) [15] for disease detection. But it failed to efficiently attain better classification outcomes. Harris hawk's optimization was developed by Ye H *et al*., (2019) [16] to distinguish the disease severity. But prediction accuracy was less. Optimized DCNN-CS Algorithm was developed by Sridevi Sakhamuri *et al*., (2022) [17] to categorize diseases. But, the performance time was not decreased. 3D Convolutional Neural Networks were developed by JV Vardhan *et al*., (2021) [18] to achieve earlier and exact outcome of disease diagnose. However, the designed network prediction time was not minimized. A hybrid ensemble modular neural network was introduced by Patricia Melin *et al*., (2020) [19] to eradicate irrelevant features and decreasing the burden involved during prediction. A case-based reasoning framework was developd by Olaide N. Oyelade and Absalom E. Ezugwu (2020) [20] for early disease detection. However, the similarity based classification was not achieved.

Incorporation of three DL models was designed in AbbasimehrH, Paki R. (2021) [21] for predicting disease. But, it failed to augment accurate disease prediction. RNN models were designed in Chandra R *et al*., (2022) [22] for disease infection prediction. A deep-LSTM ensemble model was introduced in Shastri S *et al*., (2021) [23] to diagnose disease. But, the ensemble method

was not minimizes the error rate. Recurrent and CNN schemes were introduced in Sakthivel R,ThaseenS *et al*., (2022) [24] to capture the multifaceted improvement of disease occurrences. The disease prediction model with DL was developed in Liao Z, *et al*., (2021) [25]. But, the dimensionality reduction was not achieved. DL method based on IoT was designed by Ahmed I *et al*., (2021) [26] to forecast disease. The huge volume of data was not efficiently managed. An ensemble deep learning approach was developed by Abedin MZ *et al*., (2021) [27] to predict the COVID-19 disease. However, error rate was not minimized. Another DL framework was introduced by Cui Z *et al*., (2023) [28] for disease prediction. But, higher accuracy was not attained. DL algorithm was used to learn automated diagnosis systems Mareeswari V *et al*., (2022) [29]. Time involved in the DL was not decreased. The rapid R-CNN and mask R-CNN approaches were developed by M. Emin Sahin *et al*, (2022) [25] to train and test the dataset to classify patient infections types. But, the training time was not computed.

3. Research methodology

The research methodology of proposed work is focused to improve the disease prediction performance with higher accuracy and lesser time. To perform disease prediction, medical dataset namely COVID-19 Corona Virus India Dataset and RSNA Pneumonia Detection Challenge dataset is used. COVID-19 dataset is collected from https://www.kaggle.com/datasets/imdevskp/COVID19-corona-virus-india-dataset. The COVID-19 dataset in Kaggle includes of eight.csv files such as patients data.csv, district level latest. csv, nation level daily. csv, State level daily.csv, State level latest.csv, tests day wise. csv, tests state wise. csv and Complete.csv. The COVID-19 dataset includes 262841 instances and number of features such as date, name of state / UT, latitude, longitude, total confirmed cases, death, cured/discharged/migrated, new cases, new deaths, and new recovered. Pneumonia dataset is gathered from https://github.com/LorenzG/pneumonia-kaggle/tree/master/data. The Pneumonia dataset contains dcm_data.csv, metadata.csv, stage 1 detailed class info.csv, stage 1 sample submission.csv, stage 1 train labels.csv and train positive.json. It includes 25684 instances and number of features such as patientId, boxes, Target, class, Lung Opacity, No Lung Opacity / Not Normal, Normal, Lung Opacity, No Lung Opacity / Not Normal, and Normal. The overall work is performed using hardware configurations such as Windows 10 Operating system, Intel Core I3

processor, 8GB RAM, 1TB (1000 GB) Hard disk. The architecture of proposed methodology is illustrated in Figure 1.

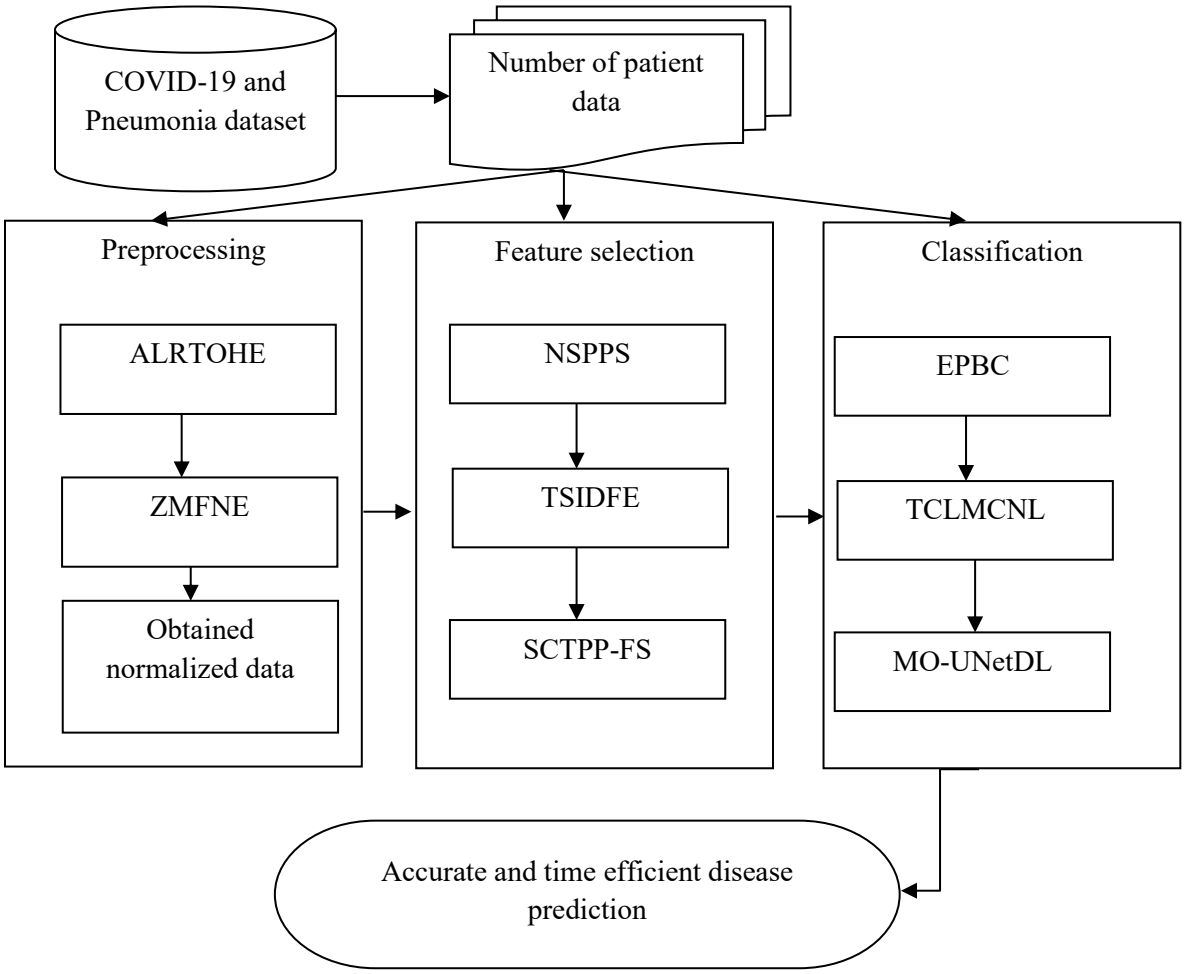

Figure 1 Overall Architecture of Proposed Methodology

Figure 1 depicts the overall design of proposed framework for disease prediction. As given in above figure, the number of patient data from the given dataset is collected as input. Upon collecting the input data, the preprocessing, feature selection and classification is carried out to

classify the data into normal or abnormal for accurate disease prediction. The above said process is briefly explained as follows.

3.1 Preprocessing

The main goal of data preprocessing is to improve the quality of the data and to make it more suitable for further processing. The proposed preprocessing phase includes two techniques Additive Log Ratio Transformed One Hot Encoding (ALRTOHE) Technique and ZMFNE technique to reduce the complexity involved in the data processing.

3.1.1 Additive Log Ratio Transformed One Hot Encoding (ALRTOHE) Technique

ALRTOHE Technique is proposed to carry out the data preprocessing with minimum time and maximum accuracy. To start with this, patient data from the given dataset is collected as input. Then, the data normalization is performed to normalize the data specific range by using additive log-ratio transformation. In second, data decoding is carried out to transfer the numerical data within binary coding. Here, the data decoding is accomplished for modifying numerical data within binary coding. From that, the binary representation of data is attained in data preprocessing with minimum time. The Pseudocode ALRTOHE algorithm is described as given below.

Input: Dataset D, features $F = f_1 f_2, .., f_m$'
Output: Obtain the pre-processed data
Begin **1**: Collect the number of features from the dataset'D' **2**: **For** each feature in the dataset 'f_i' **3.** Construct the feature matrix *'A'* **4.** Measure the log ratio with mean and standard deviation **'**LR_a**'** **5**. Normalize the data in the ranges from 0 to 1 **6**. **End for** **7.** **Apply** o*ne hot encoding* to a normalized data **8.** **Return** (binary representation of data '0' or '1') **9. Obtain the preprocessed dataset** **End**

Algorithm 1 Additive Log Ratio transformed One Hot Encoding based Data Preprocessing

Algorithm 1 illustrates the procedure of data preprocessing by using ALRTOHE for decreasing the time and space complexity metrics. Number of features and raw data is gathered as input from given dataset. Followed by, additive log-ratio transformation is applied to get the

normalized data. Next, OHE is applied for transforming the numerical data into binary values to decrease the complexity arises in the data handling.

3.1.2 Zero Mean Feature Normalized Encoding (ZMFNE) Technique

Another technique named as ZMFNE is proposed to significantly preprocess the input data with better accuracy and lesser time. The purpose of ZMFNE Technique is to clean data and make it suitable for classification which improves the accuracy and efficiency of the model. The ZMFNE Technique performs two different processes such as data normalization and encoding to get the preprocessed output. The Pseudocode ZMFNE algorithm is given as given below.

Input: Dataset D, attributes $A = a_1 a_2, .., a_m$'
Output: Obtain the preprocessed data
Begin **1:** Collect the number of attributes $A = a_1 a_2, .., a_m$ and data from the dataset'$D_i = d_1, d_2, .., d_n$' **2:** **For** each attribute in the dataset **3**: Form the input sample matrix 'M' **4**: Apply zero mean feature scaling using (4.7) **5:** Normalize the data in the ranges from 0 to 1 **6:** **End for** **7:** Perform one hot encoding and obtain the binary representation of data '0' or '1' **End**

Algorithm 2 Zero Mean Feature Normalized Encoding based Data Preprocessing

Algorithm 2 demonstrates ZMFNE data preprocessing by applying data normalization and encoding process. With the input of patient data, zero mean feature scaling is employed to normalize the input data. Next, one hot encoding is applied to get the binary representation of the data. From this, data pre-processed outcomes are achieved to diminish the time and space complexity involved in the disease prediction.

3.2 Feature Selection

Feature selection is carried out prior to the prediction process to increase the performance of disease diagnosis. The primary aim of feature selection is to diminish number of features through picking the most informative features. Reducing the number of features provides several benefits to the practitioner in the analysis process. Feature selection is performed by proposing three novel techniques.

3.2.1 Nonlinear Sammon Projective Pattern Selection (NSPPS) Model

NSPPS model is designed to choose the relevant patterns with lesser error for disease prediction. In NSPPS model, Sammon projection is applied to map the high-dimensional space to low dimensionality space to preserve the inter-point distance structure. The Nonlinear Sammon Projection is employed to choose the patterns observed from contagious disease by a newly discovered disease. By using sammon's mapping, inter-pattern distances are maintained and thereby decreases the error rate involved in choosing the pertinent patterns. Thus, the performance of feature selection using NSPPS model is improved in terms of error and time.

Input: Dataset 'DS', district level counts 'DLC', Patient-wise data 'PD', patient number 'PID' **Output**: Optimal and error-minimized relevant pattern selection 'RP'
1: **Initialize** time instance 't' **2**: **Initialize** rate at which susceptible population goes to confirmed cases 'α', susceptible population take onside as active people 'μ_1' **3**: **Initialize** rate at which susceptible population goes to active cases from confirmed cases 'β', rate at which confirmed cases are added to active cases '$\mu_2(t)$' **4**: **Initialize** the rate at which active cases goes to the deceased cases '$\gamma(N)$' **5**: **Begin** **6**: **For** each Dataset 'DS' with district level counts 'DLC' obtained from Patient-wise data 'PD'patient files with patient number 'PID' **7**: Formulate a mapping function 'fun' with maximizing criterion **8**: For each time instance 't' **9**: Estimate confirmed cases **10**: Estimate active cases **11**: Estimate deceased cases **12**: **End for** **13**: Formulate Sammons mapping **14**: **Return** (relevant patterns 'RP') **15**: **End for**

16: **End for**
17: **End**

Algorithm 3 Nonlinear Sammon Projective Pattern Selection

Algorithm 3 depicts the process of NSPPS model for feature selection. By using NSPPS, optimal and error-minimized pertinent patterns are selected. Nonlinear Sammon Projection is employed to choose the patterns identified from infectious disease. With the help of Sammon's mapping, inter-pattern distances are maintained and decreases the error rate in choosing the pertinent patterns.

3.2.2 Tversky Similarity-Indexed Distributive Feature Embedding (TSIDFE) Technique

TSIDFE Technique is proposed to choose the pertinent features for disease detection with maximum accuracy for disease prediction. Proposed TSIDFE technique is used to map the high-dimensional data in low-dimensional space of analogous features. In TSIDFE technique, Tversky index similarity coefficient is computed between two features. Based similarity coefficients, relevant and irrelevant features are identified for disease prediction. The relevant is considered for further process whereas the irrelevant is eliminated to minimize the time and complexity involved in the prediction process.

Input: Dataset, preprocessed features $F = f_1 f_2, .., f_m$
Output: Select relevant features

Begin
1. Collect the number of features $\{F = f_1 f_2, .., f_m\}$
2. **For each** feature 'f_i' and 'f_j'
3. Measure the similarity '$\beta\ (k_i, k_j)$'
4. **if** $(\delta = 1)$**then**
5. The feature is said to be relevant
6. **Select** relevant features
7. **else**
8. The feature is said to be irrelevant
9. **Remove ir**relevant features
10. **end if**
11. **end for**
End

Algorithm 4 Tversky Similarity Indexed Stochastic Distributive Feature Embedding Technique

The pseudocode representation of TSIDFE algorithm for feature selection based on the similarity measure is shown in algorithm 4. Features are collected from input dataset. Similarity index is computed discover the relevant features. When the similarity coefficient provides '1' as output, then the feature is said to be relevant. Otherwise, the feature is said to be irrelevant. The relevant features are selected for disease prediction. Accordingly, accurate feature selection is achieved with lesser complexity.

3.2.3 Statistical Correlative Targeted Projection Pursuit-based Feature Selection (SCTPP-FS)

SCTPP-FS Technique is proposed to effectively choose the pertinent features from the dataset to attain accurate disease prediction. The target features are projected through finding the correlation between the features. Here, correlation between the features is determined using Kaiser–Meyer–Olkin correlative projection pursuit. The Targeted projection pursuit is a type of dimensionality reduction model employed to pick the features. It allows interactively discovering extremely intricate data and finding the significant features using Kaiser–Meyer–Olkin correlation. It is a statistical measure for determining the well-matched features for disease prediction. This in turns, relevant features are selected with higher accuracy and lesser error rate.

Input: Preprocessed big dataset, number of features $F \in f_1, f_2, \ldots, f_n$
Output: Select significant features for disease prediction

Begin
1: Collect the number of features$F \in f_1, f_2, \ldots, f_n$ from pre-processed dataset
2: **For** each featuref_i and f_j
3: Determine the correlation between features '$\rho = \sum_{i=1}^{n} \sum_{j=1}^{m} C\ (f_i, f_j)$'
4: if$(\rho = 1)$**then**
5: Features are identified as relevant
6: else
7: Features are identified as irrelevant
8: end if
9: Project the target features into low-dimensional space '$M: T_f \rightarrow L$'
10: Remove the other irrelevant features
11: Return (target features)
12: End for
End

Algorithm 5 Statistical Correlative Targeted Projection Pursuit-based Feature Selection

Algorithm 5 provides feature selection using SCTPP-FS. The target features are mapped through finding the association between the features. The feature with maximum correlation is chosen as the target feature to predict the disease accuracy.

3.3 Classification

The classification of data is performed with the input of optimal and error-minimized pertinent features. The main goal of accurate and early classification is to instantly observe patients. Classification of patient data is accomplished by proposing three novel classifier as follows.

3.3.1 Emphasis Perceptron Boosting Classification (EPBC) technique

EPBC technique is proposed to classify the patient data for disease prediction with better accuracy. EPBC technique takes patient data with selected features as input. With this input, accurate classification is made with the help of boosting algorithm. The proposed EPBC is developed for improving the accuracy of disease prediction with lesser time. With this intend, initially weak classifier i.e.., perceptron binary classifiers are formed through the weighted sum. The designed classifiers partition the patterns with zero training error. The weight and feature vector objective function is formulated to get the predicted result at an early stage.

Input: Dataset 'DS', district level counts 'DLC', Patient-wise data 'PD', patient number 'PID' **Output**: Accurate classified results
1: **Initialize** relevant patterns '$RP = rp_1, rp_2, \dots, rp_n$', Weight '$W = w_1, w_2, \dots, w_n$' **Step 2**: **Begin** **3**: **For** each Dataset 'DS' with district level counts 'DLC' obtained from Patient-wise data 'PD'patient files with patient number 'PID' **4**: Evaluate weighted sum **5**: Formulate linear classifier **6**: Formulate weight aggregated linear classifier **7**: Evaluate objective function **8**: **Return** classified results 'b_i' **9**: **End for** **10**: **End**

Algorithm 6 Emphasis Perceptron Boosting Classification

As given in the above algorithm 6, weak classifier based on the perceptron function is modeled for classification. The emphasis boosting is created using weighted sum that partitions

the pattern with zero training error. Later, the objective function is generated by merging the weight and feature vector for getting the predicted result at an early stage.

3.3.2 Time-dependent Cox regressive Levenberg–Marquardt Convolutional Neural Learning (TCLMCNL) Technique

TCLMCNL Technique is proposed to classify the patient data classification for disease prediction. TCLMCNL technique comprises input layer, an output layer, and hidden layers. The input hidden layer takes number of data with selected features as input. Then, the time-dependent cox regression is applied to calculate the Cramér's phi correlation function at the hidden layer. Based on the regression results, precise classification is performed and the results are obtained at the output layer for disease prediction.

Input: selected features '$f_1, f_2, f_3, \ldots . f_k$', Patient data samples $D = D_1, D_2, \ldots, D_n$
Output: Increase prediction accuracy

Begin
1. Collect the relevant features $f_1, f_2, f_3, \ldots . f_k$ in input layer
2: **For** each testing feature 'f_i' **--hidden layer**
3: **For** each training feature 'f_k'
4: Apply Time-Dependent Cox regression
5: **If** $(R = 1)$**then**
6: **patient D**ata sample is classified as a diseased
7: **else**
8: **patient D**ata sample is classified as a normal
9: **end if**
10: **end for**
11: **end for**
12: **For each time step *'t'***
13: Measure Huber loss 'L_h'
14: Apply –the Levenberg Marquardt algorithm to find the minimum loss
15:*if* (arg min L_h) **then**
16: Obtain accurate classification results at the output layer
17: **else**
18: Repeat step 2
19: **end if**
20: End for
21: End for
End

Algorithm 7 Time-dependent Cox regressive Levenberg–Marquardt Convolutional Neural Learning classifier

Algorithm 7 describes the process of disease detection using TCLMCNL classifier. The classification is performed by means of examining the testing disease patterns and training data via time-dependent Cox regression. According to the time-dependent Cox regression, Cramér's phi correlation is used to obtain the classification result. Besides, Huber loss is determined and Levenberg–Marquardt algorithm is employed to get the error minimized classification results for disease prediction.

3.3.3. Memetic Optimized U-Net Deep Learning (MO-UNetDL) classifier

MO-UNetDL classifier is proposed for disease prediction via classifying patient data. MO-UNetDL technique comprises several layers to observe the patient samples. The data in the dataset is partitioned into testing and training data from the dataset. After that, the training of deep U-Net learning model is employed to classify the patient data into a normal or diseased. In the classification process, the hyperparameter of the deep learning model is optimized using Memetic algorithm get higher accuracy of the classification with lower error. With this, the accurate patient data classification outcomes are acquired.

Input: Selected relevant features $f_k = f_1, f_2, \dots, f_k$ and training data samples $D = D_1, D_2, \dots, D_n$
Output: Increase the disease prediction accuracy

Begin
1: Collect the selected features $f_k = f_1, f_2, \dots, f_k$ with data samples
2: Apply k fold cross-validation for splitting the data sent into testing and training
3: Training data are given as input to the input layer of the U-Net deep learning classifier
4: **For each training data–[hidden layer]**
5: **Perform the convolution process** with weight 'α_i' and add bias 'g'
6: **For** each training data with testing disease data
8: Measure the Wilcox's index coefficient 'Q'
9: Apply soft step activation function 'A'
10: Perform a max-pooling operation
11: Select the best value and remove others
12:If ($A = 1$) **then**
13: Correctly predicted as disease

14: **Else**
15: Correctly predicted as normal
16: End if
17: Perform upsampling process for predicting all the data samples
18: **For each classification results**
19: Measure the cross entropy loss 'LL'
20: Update the weight '∇w'
21: Find minimum loss by tuning hyperparameters using memetic optimization
22: Initialize the populations of hyperparameters such as weight, epochs, and learning rate
23: For each individual in the population
24: Calculate the fitness 'F'
25: **If** $(arg\min LL)$**then**
26: Selects an individual from the population
27: End if
28: Select the individual from the population
29: Generate new offspring using a two-point operator
30: Perform bit flip mutation for inverting the bit
31: Replace the old individual with a new one
32: Go to step 24
33: Terminate the algorithm until the optimal solution is obtained
34: **Return (Optimal hyperparameters)**
35: End for
36: End

Algorithm 8 Memetic Optimized U-Net Deep Learning classifier

As described in the above algorithm, MO-UNetDL classifier is used for accuracy disease prediction. Wilcox's index coefficient is used to discover the similarity between the features. Accordingly, soft step activation correctly diagnoses the diseased data and normal data. Also, the cross entropy loss is computed based on the actual and predicted output results. In addition, the memetic optimization algorithm is applied to tune the hyperparameters of classifier for achieving loss reduced classification results in the disease prediction.

4. Results and Discussion

Results of proposed preprocessing, feature selection and classification methods are analyzed through the implementation of Python programming language. Performance of proposed methods is analyzed using COVID-19 dataset and Pneumonia dataset in terms of accuracy, time, precision, recall, specificity and so on. Accuracy is described as the ratio of the number of data samples are accurately predicted as normal or disease to the total patient data. Time consumption

is measured as the amount of time used to identify the disease. Precision is calculated based on the true positives and false negatives. Recall is determined depended on the measure of true positives as well as false negatives of disease prediction. Specificity is determined on the basis of true negatives as well as false positives of disease prediction.

Table 1 Results of Preprocessing Techniques

Methods/Metrics	COVID-19 dataset					Pneumonia dataset				
	Training PA (%)	Testing PA (%)	Training PT (ms)	Testing PT (ms)	ER (%)	Training PA (%)	Testing PA (%)	Training PT (ms)	Testing PT (ms)	ER (%)
Existing VOC-DL	70.99	68.54	6270	610	29.00	73.72	69.25	6280	510	31.21
Existing CNN-GRU based hybrid deep learning model	75.19	72.62	5926	580	24.80	78.91	74.62	5950	460	27.42
Proposed ALRTOHE	82.43	80.36	5430	540	17.56	87.81	81.25	5450	365	19.17
Proposed ZMFNE	85.86	83.12	5060	501	14.13	89.83	84.25	5090	340	16.23

Table 1 explains the comparative performance analysis of preprocessing techniques for COVID-19 and Pneumonia datasets. By observing table values, proposed ZMFNE performed better than the other methods. Proposed ZMFNE achieved 85.86% of Preprocessing Accuracy (PA), 5060 ms of Preprocessing Time (PT), and 14.13% Error Rate (ER) in the training process for COVID-19 dataset. Also, ZMFNE achieved 89.83% of preprocessing accuracy, 5090 ms of preprocessing time, and 16.23% error rate in the training process for Pneumonia dataset.

Table 2 Results of Feature Selection Techniques

Methods/ Metrics	COVID-19 dataset					Pneumonia dataset				
	Training FA (%)	Testing FA (%)	Training FT (ms)	Testing FT (ms)	ER (%)	Training FA (%)	Testing FA (%)	Training FT (ms)	Testing FT (ms)	ER (%)
Existing Chi^2-MI	80.66	77.52	6890	675	19.33	84.67	79.52	6090	545	15.32
Existing AHEG-FS	83.64	80.36	6390	600	16.35	86.57	82.14	5480	490	13.42
Proposed NSPPS	86.16	83.54	5780	560	13.83	88.72	85.62	3990	350	11.27
Proposed TSIDFE	88.82	85.25	5315	510	11.17	91.10	87.63	3320	310	8.89
Proposed SCTPP-FS	90.79	87.14	4920	485	9.20	93.9	89.25	3070	280	6.1

Table 2 describes the results of proposed NSPPS, TSIDFE and SCTPP-FS with existing methods. As compared to other methods, proposed SCTPP-FS technique achieved better results

while selecting the significant features for disease prediction. Results of Feature selection Accuracy (FA) using SCTPP-FS is obtained as 90.79%, Feature selection Time (FT) as 4920ms, and ER as 9.20% during training using COVID-19 dataset. In addition, the FA using SCTPP-FS is obtained as 93.9%, FT as 3070 ms, and ER as 6.1% during testing process for Pneumonia dataset.

Table 3 Results of Classification Techniques

Methods/ Metrics	COVID-19 dataset					Pneumonia dataset				
	Training CA (%)	Testing CA (%)	Training CT (ms)	Testing CT (ms)	Precision (%)	Training CA (%)	Testing CA (%)	Training CT (ms)	Testing CT (ms)	Precision (%
Existing Deep Learning LSTM Ensemble model	94.14	90.25	6760	640	94.51	94.26	91.24	4021	390	95.01
Existing DSPM	95.13	91.16	6270	580	95.86	96.15	92.14	3850	360	96.40
Proposed EPBC	96.18	92.65	5640	530	97.07	97.50	93.54	3425	320	98.25
Proposed TCLMCNL	96.83	93.52	5210	490	97.95	97.10	94.25	3010	280	98.38
Proposed MO-UNetDL	97.90	94.25	4740	460	98.86	98.00	96.24	2056	190	98.56

Table 3 shows the performance results of proposed and existing classification techniques. By observing above table, proposed MO-UNetDL achieved better performance for disease prediction. Results of Classification Accuracy (CA) using MO-UNetDL is obtained as 97.90%, classification time (CT) as 4740 ms, and precision as 98.86% during training using COVID-19 dataset. Also, the CA using MO-UNetDL is obtained as 98%, CT as 2056 ms, and precision as 98.56% during testing process for Pneumonia dataset.

5. Conclusion

A novel preprocessing, feature selection and classification methods are developed for accurate- disease prediction. The proposed ALRTOHE technique is designed to preprocess the given patient data from the dataset. ZMFNE technique is developed to perform data normalization and data transformation process to reduce the complexity involved in handling patient data. Followed by the data preprocessing, feature selection is performed by using NSPPS, TSIDFE technique and SCTPP-FS technique to choose the relevant features for disease prediction. In NSPPS, Sammon projection is employed to project the high-dimensional space features to low dimensionality space features. TSIDFE calculates the Tversky index similarity between features to pick the pertinent features for disease prediction. SCTPP-FS technique reduces the error occurred in the feature selection using KMO correlative projection pursuit. Lastly, the

classification is performed using EPBC, TCLMCNL and MO-UNetDL classifier where the accurate classification of patient data with selected features are made to enhance the performance of disease prediction. The experiments are conduced to verify the effectiveness of the proposed techniques using COVID-19 and Pneumonia dataset. The results indicate the proposed techniques achieved better disease detection performance than the conventional methods.

References

[1] Anil Utku, "Deep learning based hybrid prediction model for predicting the spread of COVID-19 in the world's most populous countries", Expert Systems with Applications, Elsevier, Volume 231, 30 November 2023, Pages 1-12

[2] Zhifang Liao, Yucheng Song, Shengbing Ren, Xiaomeng Song, Xiaoping Fan, Zhining Liao, "VOC-DL: Deep learning prediction model for COVID-19 based on VOC virus variants", Computer Methods and Programs in Biomedicine, Elsevier, Volume 224, September 2022, Pages 1-13

[3] Syed Javeed Pasha, and E. Syed Mohamed, "Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction", Informatics in Medicine Unlocked, Elsevier, Volume 32, 2022, Pages 1-21

[4] Samrat Kumar Dey, Khandaker Mohammad Mohi Uddin, Hafiz Md. Hasan Babu, Md. Mahbubur Rahman, Arpita Howlader, K.M. Aslam Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease", Intelligent Systems with Applications, Elsevier, Volume 16, November 2022, Pages 1-13

[5] Sourabh Shastri, Kuljeet Singh, Sachin Kumar, Paramjit Kour, Vibhakar Mansotra, "Deep-LSTM ensemble framework to forecast COVID-19: an insight to the global pandemic", International Journal of Information Technology, Springer, Volume 13, 2021, Pages 1291–1301.

[6] Devante Ayris, Maleeha Imtiaz, Kye Horbury, Blake Williams, Mitchell Blackney, Celine Shi Hui See, Syed Afaq Ali Shah, "Novel deep learning approach to model and predict the spread of COVID-19", Intelligent Systems with Applications, Volume 14, 2022, Pages 1-12.

[7] ChunyanAo, Wenyang Zhou, Lin Gaoa, Benzhi Dong, Liang Yu, "Prediction of antioxidant proteins using hybrid feature representation method and random forest", Genomics, Elsevier, Volume 112, Issue 6, 2020, Pages 4666-4674

[8] Haolin Wang, Xuhai Tan, Zhilin Huang, Bo Panc, Jie Tian, “Mining incomplete clinical data for the early assessment of Kawasaki diseasebased on feature clustering and convolutional neural networks”, Artificial Intelligence In Medicine, Elsevier, Volume 105, 2020, Pages 1-7.

[9] Hongbin Zhang, Weinan Liang, Chuanxiu Li, Qipeng Xiong, Haowei Shi, Lang Hu, Guangli Li, “DCML: Deep contrastive mutual learning for COVID-19 recognition”, Biomedical Signal Processing and Control, Elsevier, Volume 77, 2022, Pages 1-14.

[10] Andrea Manconi, Giuliano Armano, Matteo Gnocchi and Luciano Milanesi, “A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19”, Applied science, Volume 12, 2022, Pages 1-23.

[11] Malik h, Naeem A, Naqvi R.A and Loh W-K. DMFL_Net: A Federated Learning-Based Framework for the Classification of COVID-19 from Multiple Chest Diseases Using X-rays. Sensors, MDPI, Volume 23, Issue 2, 2022, pages 1-23

[12] Gautam Y. Transfer Learning for COVID-19 cases and deaths forecast using LSTM network. ISA Transactions, Elsevier, Volume 124, 2022, Pages 124:41-56

[13] Anand Motwani, Piyush Kumar Shukla, Mahesh Pawar, Manoj Kumar, Uttam Ghosh, Waleed Alnumay, Soumya Ranjan Nayak, “Enhanced framework for COVID-19 prediction with computed tomography scan images using dense convolutional neural network and novel loss function”, Computers and Electrical Engineering, Volume 105, January 2023, Pages 1-15

[14] Hanuman Verma, Saurav Mandal, Akshansh Gupta, “Temporal deep learning architecture for prediction of COVID-19 cases inIndia”, Expert Systems With Applications, Springer, Volume 195, 2022, Pages 1-17.

[15] Le D, Parvathy VS, Gupta D, Khanna A, Rodrigues J. J. P. C, Shankar K. IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification. International Journal of Machine Learning and Cybernetics, Springer, Volume 12, 2021, Pages 3235–3248.

[16] Ye H, Wu P, Zhu T, Xiao Z, ZhangX, Zheng L, *ET AL*., Diagnosing Coronavirus Disease 2019 (COVID-19): Efficient Harris Hawks-Inspired Fuzzy K-Nearest Neighbor Prediction Methods.IEEE Access, Volume 9, 2021, Pages 17787 – 17802.

[17] Sridevi Sakhamuri, Dr. K. Kiran Kumar, “Deep learning and metaheuristic algorithm for effective classification and recognition of paddy leaf diseases”, Journal of Theoretical and Applied Information Technology, Volume 100, Issue 4, 2022, Pages 1-11

[18] JV Vardhan, PC Vemula, G Srinivas, G Chowdary, SB Kumar, “Diagnosis of COVID-19 using 3d convolutional neural networks”, Journal of Theoretical and Applied Information Technology, Volume 99, Issue 24, 2021, Pages 5794-5803

[19] Patricia Melin, Julio Cesar Monica, Daniela Sanchez, Oscar Castillo, “A new prediction approach of the COVID-19 virus pandemic behaviorwith a hybrid ensemble modular nonlinear autoregressive neuralnetwork”, Soft Computing, Springer, Volume 27, Issue 5, 2020, Pages 1-10

[20] Olaide N. Oyelade, Absalom E. Ezugwu, “A case-based reasoning framework for early detection and diagnosis of novel coronavirus”, Informatics in Medicine Unlocked, Elsevier, Volume 20, 2020, Pages 1-22.

[21] AbbasimehrH, Paki R. Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization. Chaos, Solitons and Fractals, Elsevier, Volume 142, 2021, Pages1-37

[22] Chandra R, Jain A, Chauhan DS. Deep learning via LSTM models for COVID-19 infection forecasting in India. PLoS ONE, Volume 17, Issue 1, 2022, Pages 1-28

[23] Shastri S, Singh K, Kumar S, Kour P, Mansotra V. Deep-LSTM ensemble framework to forecast COVID-19: an insight to the global pandemic. International Journal of Information Technology, Springer, Volume 13, 2021, Pages 1291–130

[24] Sakthivel R,ThaseenS, Vanitha M, Deepa, Angulakshmi,Mangayarkarasi, *et al*., An efficient hardware architecture based on an ensemble of deep learningmodels for COVID -19 prediction, Electronics, Volume 9, Issue 4, 2022, Pages 1-13.

[25] Liao Z, Lan P, Fan X, Kelly B, Innes A, Liao Z. SIRVD-DL: A COVID-19 deep learningprediction model based ontime-dependent SIRVD. Computers in Biology and Medicine, Volume 138, 2021, Pages 1-15

[26] Ahmed I, Ahmad A, Jeon G. An IoT-Based Deep Learning Framework for Early Assessment of COVID-19, IEEE Internet of Things Journal, Volume 8, Issue 21, 2021, Pages15855 – 15862.

[27] Abedin MZ, Moon MH, M. Hassan K, Hajek P. Deep learning-based exchange rate predictionduring the COVID-19 pandemic., Springer, 2021, Pages1-52

[28] Cui Z, Wu J, Lian W, Wang Y. A novel deep learning framework with a COVID-19 adjustment for electricity demand forecasting. Energy Reports, Elsevier, Volume 9, 2023, Pages 1887-1895

[29] Mareeswari V, Vijayan R ,Sathiyamoorthy E and Ephzibah EP. A narrative review of medical image processing by deep learning models: origin to COVID-19. International Journal of Advanced Technology and Engineering Exploration, Volume 9, Issue 90, 2022, Pages 623-643.
[30] M. Emin Sahin, Hasan Ulutas, Esra Yuce, Mustafa Fatih Erkoc,"Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images", Neural computing and application, Volume 35, 2022, Pages13597–13611

PUBLICATIONS

International Conferences

- Presented paper on "A Study on Sentiment Analysis in CORONA Lockdown" in the International Conference on Continuity, Consistency and Innovation in Applied Sciences and Humanities at St.Martin's Engineering College, Secundrabad on 13th August 2020.
- Presented paper on "Survival Study on COVID Disease Prediction Methods with Patient Data" in the International Conference on Advanced Computing at Bharathiar University, Coimbatore on 2nd and 3rd March 2023.
- Presented paper on "A Study on COVID Disease and Deep Learning Techniques" in the International Conference on Cognitive Computing – A Disruptive Technology at Sri Krishna Arts and Science College on 18.03.2024.

Journal Publications (Scopus Indexed)

- Published phase I work paper "Early COVID-19 Spread Prediction Using Sammon Projective and Perceptron Boosting "in "Journal of Theoretical and Applied Information Technology" on 31st May 2022. Vol.100. No 10 . *ISSN: 1992-8645 , E-ISSN: 1817-3195.* http://www.jatit.org/volumes/Vol100No10/18Vol100No10.pdf
- Published phase II work paper "Additive Log Transformation Distributed Feature Embedding Convolutional Neural Learning Classifier For Early COVID-19 Prediction" in "International Journal of Advanced Technology and Engineering Exploration" on 31st May 2023, Vol 10(102), ISSN (Print): 2394-5443 ISSN (Online): 2394-7454 , https://dx.doi.org/10.19101/IJATEE.2022.10100070.